非线性常微分方程的计算不确定性原理*
——Ⅱ. 理论分析

李建平 曾庆存
(中国科学院大气物理研究所大气科学和地球流体力学数值模拟国家重点实验室, 北京 100029)
丑纪范
(北京气象学院, 北京 100081)

摘要 研究了常微分方程一般数值解法的误差传播规律, 提出理论收敛性、数值收敛性和真实收敛性 3 种收敛性概念, 详细讨论了浮点机上一般数值解法的舍入误差的各种分量. 通过引进一类新的递推不等式, 本质改进了线性多步法误差界的经典结果, 结合概率理论导出了浮点机上舍入误差的"正常"累积增长, 并给出一般多步法总误差的统一估计. 在此基础上, 解释了数值试验中的各种现象, 导得两个与方程、初值、数值格式无关且与数值试验中相一致的普适关系, 并给出计算不确定性原理的明确数学表述, 阐明了数值解法和计算机所带来的两种不确定性之间存在的固有关系.

关键词 计算不确定性原理 舍入误差 离散化误差 普适关系 机器精度

在文献[1]中, 我们给出非线性常微分方程数值求解中的一些新的数值现象, 指出由于机器精度的有限性所导致的舍入误差对非线性常微分方程数值计算的重要影响, 提出计算不确定性原理. 为了从理论上给出文献[1]中数值结果的解释和证明, 并且使所得结论具有普遍意义和广泛的使用价值, 那么就必须对常微分方程数值解法的误差进行细致研究, 必须考虑机器的有限精度性所造成的舍入误差的影响. 关于常微分方程组数值解法的误差分析已有较多研究, 其中离散化误差的经典结果可以在文献[2~10]中找到, 而舍入误差的系统研究(主要是针对定点机)则是 Henrici[2,3]作出的. 然而, 纵观这些经典结果, 除了单步法外, 对于线性多步法的误差估计是非常粗糙的, 不能为本文的分析所用. 同时, 目前也没有一般多步法的误差估计的统一公式. 而且, 对于浮点机上一般多步法的舍入误差的细致理论分析亦很稀少. 所以, 为了使本文的结果有广泛的适用性, 就必须对经典结果进行本质的改进, 并得到一般多步法的统一误差估计, 尤其是要得到浮点机上舍入误差的"正常"(或真实)累积增长. 本文通过引入一类新的递推不等式, 不仅改进了经典误差界, 而且获得了一般多步法的

2000-04-25 收稿
* 国家重点基础研究发展规划(批准号: G1998040900)、国家自然科学基金(批准号:49805006,49905007)、优秀国家重点实验室(批准号:49823002)资助项目和中国科学院资源环境领域知识创新工程重要方向项目及中国科学院大气物理研究所创新项目

统一的误差界；尤其利用概率理论给出了浮点机上舍入误差的“正常”累积增长，从而合理解释了文献[1]数值试验中的各种现象，并给出计算不确定性原理的明确数学表述．

1　基本描述

考虑如下一阶 m 维常微分方程组初值问题：

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \boldsymbol{y}' = \boldsymbol{f}(t,\boldsymbol{y}),\quad \boldsymbol{y}(t_0) = \boldsymbol{y}_0, \tag{1}$$

其中向量 $\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^{\mathrm{T}}$，$t \in [a,b]$，且 $\boldsymbol{f}(t,\boldsymbol{y}) = (f_1(t,\boldsymbol{y}), f_2(t,\boldsymbol{y}), \cdots, f_m(t,\boldsymbol{y}))^{\mathrm{T}}$是已知的连续向量函数．为了下面的讨论，总假定向量值函数 $\boldsymbol{f}(t,\boldsymbol{y})$在区域 $\boldsymbol{S} = \{(t,\boldsymbol{y}) \mid a \leqslant t \leqslant b, \boldsymbol{y} \in \mathbb{R}^m\}$上有定义且是连续的，$\boldsymbol{f}(t,\boldsymbol{y})$相对于 $\boldsymbol{y}$ 满足 Lipschitz 条件，即对于任意 $t \in [a,b]$和任意两个向量 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$，存在常数 L，使得

$$\| \boldsymbol{f}(t,\boldsymbol{y}_1) - \boldsymbol{f}(t,\boldsymbol{y}_2) \| \leqslant L \| \boldsymbol{y}_1 - \boldsymbol{y}_2 \| \tag{2}$$

成立，L 称为$\boldsymbol{f}(t,\boldsymbol{y})$对 $\boldsymbol{y}$ 的 Lipschitz 常数．这样初值问题(1)存在惟一连续可微的解 $\boldsymbol{y}(t)$．

众所周知，求解一阶常微分方程组初值问题的方法和结果本质上与未知函数的个数 m 无关[5]．因此以下我们常常在形式上限定只考虑一阶的、只含一个未知函数的单个常微分方程(即 $m=1$)．然而，只要诸如量 $y, f(t,y), \Phi, \Delta, e(t;h), r(t;h), E(t;h), T_k(t;y;h), \tau_k(t;y;h), \varepsilon(t;h), R$ 和 z 等等被解释为向量，这些结果照例对方程组也是正确的．为了不失一般性，在适当的地方我们仍然用范数 $\|\cdot\|$ 代替 $|\cdot|$．因此，所有取范数 $\|\cdot\|$ 的量都被当作向量．

一个一般的数值方法可用统一的公式表示[6]：

$$\begin{cases} \alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = h\Phi(t_n, y_{n+k}, y_{n+k-1}, \cdots, y_n; h; f), & 0 \leqslant n \leqslant N-k; \\ y_j = y(t_0 + jh), & 0 \leqslant j \leqslant k-1, \end{cases} \tag{3}$$

其中 $\alpha_j (j = 0,1,\cdots,k)$是不依赖于 n 的实常数，$\alpha_k \neq 0$，k 是一个固定的正整数，h 是步长，$y_0, y_1, \cdots, y_{k-1}$已知且 $y_j = y(t_j; h)$ $(j = 0, 1, \cdots, N)$．方程(3)称为一般的 k 步法．它包含了所有通常的数值方法作为特例．当 $k=1$ 则得一单步方法，此时 Φ 称为方法的增量函数．如果 $k>1$ 则得一多步方法．如果函数 Φ 与 y_{n+k}无关，那么(3)式称为是显式方法；否则，它是一个隐式方法．对于 Φ 的不同选择将得到不同的数值方法．为简单起见 Φ 中的自变量f 将被略去．为讨论方便，定义

$$\rho_k(\xi) = \alpha_k \xi^k + \alpha_{k-1}\xi^{k-1} + \cdots + \alpha_0, \tag{4}$$

称为 k 步法(3)的特征多项式．在(3)式中取 $k=1$，可得一般的显式单步方法

$$y_{n+1} = y_n + h\Phi(t_n, y_n; h). \tag{5}$$

Taylor 级数法和 Runge-Kutta 法是两种著名的单步法．在(3)式中令

$$\Phi(t_n, y_{n+k}, y_{n+k-1}, \cdots, y_n; h) = \beta_k f(t_{n+k}, y_{n+k}) + \beta_{k-1} f(t_{n+k-1}, y_{n+k-1}) + \cdots + \beta_0 f(t_n, y_n),$$

则方法(3)成为

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y_{n+j}). \tag{6}$$

由于函数 Φ 线性地依赖于 f，所以(6)式称为线性多步法或更准确地说是一个线性 k 步法. (6)式当 $\beta_k=0$ 是显式的，当 $\beta_k\neq 0$ 是隐式的. 多项式

$$\sigma_k(\xi)=\beta_k\xi^k+\beta_{k-1}\xi^{k-1}+\cdots+\beta_0 \tag{7}$$

也称为 k 步法(7)的特征多项式. 显式和隐式 Adams 方法是线性多步法的两个重要特例.

求解一个微分方程或微分方程组的数值方法的真实误差有两个基本的来源：一个是近似的方法，即数值方法不能产生给定微分方程的准确解(即使计算没有舍入误差)，

$$e(t;h)=y(t)-y(t;h) \tag{8}$$

称为全局离散化(或截断)误差，其中 $y(t;h)$ 代表在给定步长 h 下数值方法给出的准确解，也叫做理论近似值. 这个误差依赖于给定的初值问题、所用数值方法、步长 h 和步数 n(即 t). 另一个误差源是由于实际计算机的有限精度导致在实际中 $y(t;h)$ 不能被准确的计算. 令 $\tilde{y}(t;h)$ 代表 $y(t;h)$ 的真实计算值并称为数值近似值，则

$$r(t;h)=y(t;h)-\tilde{y}(t;h) \tag{9}$$

称为累积的(或全局的)舍入误差. 这个误差不仅依赖于给定的微分方程和数值方法，也依赖于所用计算机、定点或浮点运算、机器所用的数字系统、计算程序的细节，特别是依赖于机器精度. 记总误差

$$E(t;h)=y(t)-\tilde{y}(t,h)=e(t;h)+r(t;h), \tag{10}$$

利用三角不等式知

$$\|E(t;h)\|\leqslant\|e(t;h)\|+\|r(t;h)\|. \tag{11}$$

定义 1[5] k 步法(3)式在点 $t_{n+k}=a+(n+k)h$ 处的(绝对)局部(离散化或截断)误差定义为

$$y(t_{n+k})-y_{n+k}, \tag{12}$$

其中 $y(t_{n+k})$ 是方程(3)的准确解，y_{n+k} 是由(3)式用 k 个准确的起始值得到的准确数值解 $y_{n+j}=y(t_{n+j};h)(j=0,1,\cdots,k-1)$.

关于局部误差的一个实用定义如下：

定义 2 令 $y(t)$ 是方程(3)的准确解，则

$$T_k(t,y;h)=\sum_{j=0}^{k}\alpha_j y(t+jh)-h\Phi_k(t,y;h), \tag{13}$$

其中 $\Phi_k(t,y;h)=\Phi(t,y(t+kh),\cdots,y(t);h)$ 及

$$\tau_k(t,y;h)=\frac{1}{h}T_k(t,y;h) \tag{14}$$

分别称为 k 步法(3)在点 $(t+kh,y)$ 的(绝对)局部离散化(或截断)误差和相对的局部离散化(或截断)误差(相对于步长 h).

注 对显式单步法，$\Phi(t,y;h)=\Phi_1(t,y;h)$.

引理 1 考虑微分方程(1)，令 $y(t)$ 是它的准确解，Φ 是连续可微的函数. 对于局部误差(14)式有

$$y(t_{n+k})-y_{n+k}=\left(\alpha_k I-h\frac{\partial}{\partial y}\Phi(t_n,\bar{y}_{n+k},y(t_{n+k-1}),\cdots,y(t_n);h)\right)^{-1}T_k(t,y;h), \tag{15}$$

其中 I 是一个单位矩阵. 这里如果 Φ 是一个标量函数，则 $\bar{y}_{n+k}$ 是介于 $y(t_{n+k})$ 和 y_{n+k} 间的一

个值．如果 Φ 是一个向量函数，则 $\frac{\partial}{\partial y}\Phi(t_n,\bar{y}_{n+k},y(t_{n+k-1}),\cdots,y(t_n);h)$ 是 Jacobi 矩阵．

这个引理表明 $T_k(t,y;h)$ 本质上等价于局部误差．所以，定义 2 和定义 1 是等价的．

定义 3　由(3)式给出的方法称为是理论收敛的，如果对任意固定的 $t\in[a,b]$，$t=a+nh=t_n$，

$$\lim_{h\to 0}e(t;h)=0,\ \text{或}\ \lim_{h\to 0}\|y(t)-y(t;h)\|=0. \tag{16}$$

这里的理论收敛性也就是通常所讲的数值方法收敛性．理论收敛性仅仅能保证当 $h\to 0$ 理论近似值 $y(t;h)$ 将任意好的逼近准确解 $y(t)$，但不能保证随着 $h\to 0$ 数值解 $\tilde{y}(t;h)$ 收敛到 $y(t)$．因此，我们给出真实收敛性和数值收敛性的概念．

定义 4　对任意固定的 $t\in[a,b]$，$t=a+nh=t_n$，如果

$$\lim_{h\to 0}E(t;h)=0,\ \text{或}\ \lim_{h\to 0}\|y(t)-\tilde{y}(t;h)\|=0, \tag{17}$$

则由(3)式给出的方法称为是真实收敛的；如果

$$\lim_{h\to 0}r(t;h)=0,\ \text{或}\ \lim_{h\to 0}\|y(t;h)-\tilde{y}(t;h)\|=0, \tag{18}$$

则由(3)式给出的方法称为是数值收敛的．

显然，理论收敛性(16)式并不保证真实收敛性(17)式，反之亦然．一个数值方法只有当它不仅是理论收敛的而且也是数值收敛的时候才是真实收敛的．反之，根据定义一般地说真实收敛性既不蕴含理论收敛性也不蕴含数值收敛性．

为了研究舍入误差需要引入如下的局部舍入误差的概念：

定义 5　k 步法(3)的(绝对)局部舍入误差定义为

$$\tilde{y}(t)=y(t)+\varepsilon(t;h),\quad t=t_0+jh,\quad j=0,1,\cdots,k-1,$$

$$\sum_{j=0}^{k}\alpha_j\tilde{y}(t+jh;h)=h\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h)+\varepsilon(t;h),\quad n=0,1,\cdots, \tag{19}$$

其中 $\tilde{y}(t;h)$ 是准确解 $y(t)$ 的理论计算值 $y(t;h)$ 的数值近似值．

$$\delta(t;h)=\frac{\varepsilon(t;h)}{h} \tag{20}$$

称为方法的相对局部舍入误差(相对于步长 h)．

2　改进的离散化误差先验界

2.1　线性多步法

在具体讨论之前，先介绍一下有关线性多步法离散化误差先验估计的经典结果．为此，引入 Henrici[2,3] 得到的两个引理，它们在后面误差估计的改进及描述中仍然需要．

引理 2[2,3]　令特征多项式 $\rho_k(\xi)$ 满足根条件，且系数 $\gamma_j(j=0,1,\cdots)$ 定义为

$$\frac{1}{\alpha_k+\alpha_{k-1}\xi+\cdots+\alpha_0\xi^k}=\gamma_0+\gamma_1\xi+\gamma_2\xi^2+\cdots, \tag{21}$$

则

$$\Gamma=\sup_{j=0,1,\cdots}|\gamma_j|<\infty.$$

引理 3[2,3]　考虑非齐次线性微分方程

$$\alpha_k z_{m+k} + \alpha_{k-1} z_{m+k-1} + \cdots + \alpha_0 z_m =$$
$$h(\beta_{k,m} z_{m+k} + \beta_{k-1,m} z_{m+k-1} + \cdots + \beta_{0,m} z_m) + \lambda_m, \tag{22}$$

令特征多项式 $\rho_k(\xi)$满足根条件，且令

$$\sum_{j=0}^{k} |\beta_{j,n}| \leqslant \beta^*, \quad |\beta_{k,n}| \leqslant \beta, \quad \|\lambda_n\| \leqslant \Lambda, \quad n = 0,1,\cdots,N, \tag{23}$$

其中 B^*, β, Λ 是常数，再令 $0 \leqslant h < |\alpha_k| \beta^{-1}$，则(22)式的每个解当

$$\|z_j\| \leqslant z_{(0)} \quad (j = 0,1,\cdots,k-1)$$

时满足

$$\|z_n\| \leqslant \Gamma^* (Akz_{(0)} + n\Lambda) e^{nhL^*}, \quad n = 0, 1, \cdots, N, \tag{24}$$

其中

$$L^* = \Gamma^* B^*, \quad \Gamma^* = \frac{\Gamma}{1 - h\beta |\alpha_k|^{-1}}, \quad A = |\alpha_k| + |\alpha_{k-1}| + \cdots + |\alpha_0|. \tag{25}$$

对于线性多步法(6)，利用上述引理，有如下经典结果：

定理 1[2,3] 令函数 $f(t,y)$满足 Lipschitz 条件，且相对局部离散化误差

$$\|\tau_k(t,y;h)\| \leqslant \tau(h), \quad t \in [a,b], \ h \leqslant h_0, \tag{26}$$

其中 $\tau(h)$仅仅依赖于 h 和一些常数，若(6)式满足相容性条件和根条件，则对于 $h|\alpha_k^{-1}\beta_k| L < 1$ 和 $t \in [t_0, b]$，线性 k 步法的全局离散化误差

$$\|e(t;h)\| \leqslant \Gamma^* [Ake_{(0)} + (t - t_0)\tau(h)] e^{L\Gamma^* B(t-t_0)}, \tag{27}$$

其中 $e_{(0)}$是由 $e_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} \|e_j\|$ 定义的最大初始值误差，且

$$A = \sum_{j=0}^{k} |\alpha_j|, \quad B = \sum_{j=0}^{k} |\beta_j|, \quad \Gamma^* = \frac{\Gamma}{1 - h|\alpha_k^{-1}\beta_k| L}, \quad t - t_0 = nh, \quad a \leqslant t_0.$$

因为项$(t - t_0)\tau(h)$依赖于变量 t，所以界(27)式是十分粗糙的，并且不能被本文的分析所利用. 现在我们来改进它. 像单步法中的离散化误差界一样，我们希望对线性多步法上面提到的项只与一些常数有关而与变量 t 无关. 而且，我们期望能给出一般的 k 步法(3)的误差界(不局限于线性多步法). 为此，先引入如下关键的引理———一类新的递推不等式：

引理 4[1] 如果数列 ξ_n 满足如下形式的不等式：

$$|\xi_n| \leqslant A \sum_{j=0}^{n-1} |\xi_j| + mB + C\xi_{(0)}, \quad n = k, k+1, \cdots, \tag{28}$$

其中 A, B 和 C 是与 n 无关的非负常数，k 是自然数，$m = n - K, K \leqslant k$ 是整数，且 $\xi_{(0)} =$

---

1) 更一般地有如下结果：如果数列 ξ_n 满足如下不等式：

$$|\xi_n| \leqslant A \sum_{j=0}^{n-1} |\xi_j| + (n - k + 1)B + C\xi_{(0)}, \quad n = k, k+1, \cdots,$$

其中 $A \geqslant -1$, $B \geqslant 0$ 和 $C \geqslant 0$ 是与 n 无关的常数，k 是自然数且 $\xi_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} |\xi_j|$，则

$$|\xi_n| \leqslant (1 + A)^n N_1(C)\xi_{(0)} + \begin{cases} \dfrac{B}{A}[(1 + A)^{n-k+1} - 1], & A \neq 0; \\ (n - k + 1)B, & A = 0 \end{cases}$$

对 $n = k, k+1, \cdots$成立，其中 $N_1(C)$由(30)式定义.

$\max\limits_{0\leqslant j\leqslant k-1}|\xi_j|$，那么

$$|\xi_n|\leqslant(1+A)^nN_1(C)\xi_{(0)}+\begin{cases}\dfrac{B}{A}[(1+A)^m-1], & A>0;\\ mB, & A=0\end{cases}\tag{29}$$

对 $n=N_0(K),N_0(K)+1,\cdots$成立．这里函数 $N_a(x)$定义为

$$N_a(x)=\begin{cases}a, & x\leqslant a;\\ x, & x>a.\end{cases}\tag{30}$$

证　(30)式蕴含

$$N_a(x)\geqslant a,\tag{31}$$

$$N_a(x)\geqslant x,\tag{32}$$

所以对于 $A=0$ 时(29)式成立．

对于 $A>0$，用归纳法证明．如果 $A>0$，对 $j=0,1,\cdots,k-1$,因为$|\xi_j|\leqslant\xi_{(0)}$,$n=N_0(K)$,$K\leqslant k$,则 $m=N_0(K)-K$,$N_0(K)\leqslant k$,从(31)和(32)式可得

$$\begin{aligned}|\xi_{N_0(K)}|&\leqslant A\sum_{j=0}^{N_0(K)-1}|\xi_j|+mB+C\xi_{(0)}\\&\leqslant(1+AN_0(K))N_1(C)\xi_{(0)}+(N_0(K)-K)B.\end{aligned}$$

利用如下事实：

$$(1+x)^k\geqslant 1+kx,\tag{33}$$

其中实数 $x\geqslant 0$，k 是非负数，且

$$k\leqslant\frac{(1+x)^k-1}{x},\tag{34}$$

对任意的 $x>0$ 和非负数 k 成立，我们有

$$|\xi_{N_0(K)}|\leqslant(1+A)^{N_0(K)}N_1(C)\xi_{(0)}+\frac{B}{A}[(1+A)^m-1].$$

现假定(29)式对于 $n(n\geqslant N_0(K))$成立，把它代到(28)式的右端，并利用(31)～(34)式，有

$$\begin{aligned}|\xi_{n+1}|&\leqslant A\sum_{j=0}^{N_0(K)-1}|\xi_n|+A\Big\{N_1(C)\xi_{(0)}\sum_{j=N_0(K)}^{n}(1+A)^j+\frac{B}{A}\sum_{j=N_0(K)}^{n}[(1+A)^{j-K}-1]\Big\}\\&\quad+(m+1)B+N_1(C)\xi_{(0)}\\&\leqslant N_1(C)\xi_{(0)}\Big[(1+AN_0(K))+A\sum_{j=N_0(K)}^{n}(1+A)^j\Big]+B\Big[(N_0(K)-K)\\&\quad+\sum_{j=N_0(K)-K}^{m}(1+A)^j\Big]\\&\leqslant N_1(C)\xi_{(0)}\Big[(1+A)^{N_0(K)}+A\sum_{j=N_0(K)}^{n}(1+A)^j\Big]\\&\quad+\frac{B}{A}\Big[(1+A)^{N_0(K)-K}+A\sum_{j=N_0(K)-K}^{m}(1+A)^j-1\Big].\end{aligned}$$

应用恒等式

$$(1+x)^k + x\sum_{j=k}^{n}(1+x)^j = (1+x)^{n+1}, \tag{35}$$

其中 x 是任意的实数，k 是非负整数，有

$$|\xi_{n+1}| \leqslant (1+A)^{n+1}N_1(C)\xi_{(0)} + \frac{B}{A}[(1+A)^{m+1}-1],$$

因此(29)式对于 $n+1$ 成立. 引理由归纳法证得. 证毕.

利用不等式 $1+x \leqslant e^x$，其中 x 是任意的，(29)式可写成如下形式：

$$|\xi_n| \leqslant N_1(C)\xi_{(0)}e^{nA} + \begin{cases}\dfrac{B}{A}(e^{mA}-1), & A \neq 0;\\ mB, & A = 0,\end{cases} \tag{36}$$

其中 $A \geqslant 0, B \geqslant 0$.

由引理4可得

引理5 在引理3的条件下，(22)式的每个解满足

$$\|z_n\| \leqslant N_1(\eta)z_{(0)}e^{nhL^*} + \frac{\Lambda}{hB^*}(e^{nhL^*}-1), \quad n = 0, 1, \cdots, N, \tag{37}$$

其中 $\eta = A\Gamma^* k$，$A = |\alpha_{k-1}| + \cdots + |\alpha_0|$，其他常数同引理3.

证 首先当 l 为负整数时，令 $\gamma_l = 0$，则从(21)式可得等式

$$\alpha_k\gamma_l + \alpha_{k-1}\gamma_{l-1} + \cdots + \alpha_0\gamma_{l-k} = \begin{cases}1, & l = 0,\\ 0, & l \geqslant 0.\end{cases} \tag{38}$$

对固定的 n 和 $l = 0, 1, \cdots, n-k$，用由(21)式定义的 γ_l 乘相应于 $m = n-k-l$ 的方程(22)并把它们相加，在左边得到

$$\begin{aligned}&(\alpha_k z_n + \alpha_{k-1}z_{n-1} + \cdots + \alpha_0 z_{n-k})\gamma_0 + (\alpha_k z_{n-1} + \alpha_{k-1}z_{n-2} + \cdots + \alpha_0 z_{n-k-1})\gamma_1 + \cdots\\ &\quad + (\alpha_k z_{n-j} + \alpha_{k-1}z_{n-1-j} + \cdots + \alpha_0 z_{n-k-j})\gamma_{n-k} + \cdots + (\alpha_k z_k + \alpha_{k-1}z_{k-1} + \cdots + \alpha_0 z_0)\gamma_{n-k}\\ &= \alpha_k\gamma_0 z_n + (\alpha_k\gamma_1 + \alpha_{k-1}\gamma_0)z_{n-1} + \cdots + (\alpha_k\gamma_{n-k} + \alpha_{k-1}\gamma_{n-k-1} + \cdots + \alpha_0\gamma_{n-2k})z_k\\ &\quad + (\alpha_{k-1}\gamma_{n-k} + \alpha_{k-2}\gamma_{n-k+1} + \cdots + \alpha_0\gamma_{n-2k+1})z_{k-1} + \cdots + \alpha_0\gamma_{n-k}z_0\\ &= z_n + (\alpha_{k-1}\gamma_{n-k} + \alpha_{k-2}\gamma_{n-k-1} + \cdots + \alpha_0\gamma_{n-2k+1})z_{k-1} + \cdots + \alpha_0\gamma_{n-k}z_0,\end{aligned}$$

在右边有

$$\begin{aligned}&h[\beta_{k,n-k}\gamma_0 z_n + (\beta_{k-1,n-k}\gamma_0 + \beta_{k,n-k-1}\gamma_1)z_{n-1} + \cdots + (\beta_{0,n-k}\gamma_0 + \cdots + \beta_{k,n-2k}\gamma_k)z_{n-k}\\ &\qquad + \cdots + \beta_{0,0}\gamma_{n-k}z_0] + \lambda_{n-k}\gamma_0 + \lambda_{n-k-1}\gamma_1 + \cdots + \lambda_0\gamma_{n-k},\end{aligned}$$

取范数，并利用(23)和(25)式，我们有

$$\|z_n\| \leqslant h\beta|\alpha_k|^{-1}\|z_n\| + h\Gamma B^*\sum_{j=0}^{n-1}\|z_j\| + (n-k+1)\Gamma\Lambda + A\Gamma k z_{(0)}.$$

对 $\|z_n\|$ 求解，得到

$$\|z_n\| \leqslant hL^*\sum_{j=0}^{n-1}\|z_j\| + n\Gamma^*\Lambda + A\Gamma^* k z_{(0)}.$$

根据引理4，立即有

$$\|z_n\| \leqslant N_1(\eta)z_{(0)}e^{nhL^*} + \frac{\Lambda}{hB^*}(e^{nhL^*}-1), \quad n = 0, 1, \cdots, N.$$

证毕．

这个引理指出界(24)式已经被本质地改进，因为在界(37)式中的项 $\Lambda/(hB^*)$ 是与变量 $t_n = t_0 + nh$ 无关的，它取代了界(24)式中的项 $n\Lambda$ ($= (t_n - t_0)/h$). 界(37)式还可改进一点. 根据引理 5 的证明知

$$z_n = \sum_{j=0}^{k-1} A_{j,n} z_j + h \sum_{j=0}^{n} B_{j,n} z_j + \sum_{j=0}^{n-k} \lambda_j \gamma_{n-k-j}, \tag{39}$$

其中

$$A_{J,n} = \sum_{j=0}^{J} \alpha_{J-j} \gamma_{n-k-j}, \qquad J = 0, 1, \cdots, k-1, \quad k \leqslant n \leqslant N,$$

$$B_{J,n} = \begin{cases} \sum_{j=0}^{J} \beta_{k-J+j,\, n-k-j} \gamma_j, & J = 0, 1, \cdots, k-1; \\ \sum_{j=0}^{k} \beta_{j,n-J-j} \gamma_{J-k+j}, & J = k+1, k+2, \cdots, n-k, \quad k \leqslant n \leqslant N, \\ \sum_{j=0}^{n-J} \beta_{j,n-J+j} \gamma_{J-k-j}, & J = n-k+1, n-k+2, \cdots, n, \end{cases} \tag{40}$$

令

$$a = \max_{\substack{0 \leqslant J \leqslant k-1 \\ k \leqslant n \leqslant N}} |A_{J,n}|, \quad b = \max_{\substack{0 \leqslant J \leqslant N \\ k \leqslant n \leqslant N}} |B_{J,n}|, \quad \beta = \max_{0 \leqslant n \leqslant N} |\beta_{k,n}|, \tag{41}$$

则有

$$\| z_n \| \leqslant h\beta |\alpha_k|^{-1} \| z_n \| + hb \sum_{j=0}^{n} z_j + (n-k+1)\Lambda\Gamma + akz_{(0)},$$

其中 $z_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} \| z_j \|$. 由引理 4，有

引理 6　令特征多项式 $\rho_k(\xi)$ 满足根条件，并令 $0 \leqslant h < \beta^{-1}|\alpha_k|$，则(22)式的每个解满足

$$\| z_n \| \leqslant N_1(\eta) z_{(0)} e^{nh\dot{L}} + \frac{\Lambda\Gamma}{hb}(e^{nhb/c} - 1), \quad n = 0, 1, \cdots, N, \tag{42}$$

其中 $\eta = ak/c$，$c = 1 - h\beta|\alpha_k|^{-1}$，$a$，$b$ 和 β 由(41)式给出.

进一步，令

$$d = \max_{0 \leqslant n \leqslant N} (\alpha_k^{-1} \beta_{k,n}), \tag{43}$$

可证明

引理 7　令特征多项式 $\rho_k(\xi)$ 满足根条件，且 $hd < 1$ 和 $h \geqslant 0$，则(22)式的每个解满足

$$\| z_n \| \leqslant N_1(\eta) z_{(0)} e^{nh\dot{L}} + \frac{\Lambda\Gamma}{hb}(e^{nhb/c^*} - 1), \quad n = 0, 1, \cdots, N, \tag{44}$$

其中 $\eta = ak/c^*$，$c^* = |1 - dh|$，a 和 b 由(41)式给出，d 由(43)式定义.

在上述引理的基础上，现在来对线性多步法离散化误差的估计(27)式进行本质的改进.

定理 2　在定理 1 的条件下，如果 $h|\alpha_k^{-1}\beta_k|L < 1$，$t \in [t_0, b]$，那么对于 $h \leqslant h_0$，$t - t_0 = nh$，$n = 0, 1, \cdots$，线性 k 步法(6)的全局离散化误差满足

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} e^{LB\Gamma^*(t-t_0)} + \tau(h) \frac{e^{LB\Gamma^*(t-t_0)} - 1}{BL}, \tag{45}$$

其中 $e_0 = \max\limits_{0\leqslant j\leqslant k-1} \| e_j \|$, $\eta = A\Gamma^* k$, $A = |\alpha_{k-1}| + \cdots + |\alpha_0|$, 其他常数同定理 1.

证 从精确值 $y(t_{n+j})$满足的关系式

$$\sum_{j=0}^{k} \alpha_j y(t_{n+j}) = h\sum_{j=0}^{k} \beta_j f(t_{n+j}, y(t_{n+j})) + h\tau_k(t_n, y(t_n); h)$$

减去线性 k 步法(6)式. 记 $e_j = y(t_j) - y_j, j = 0,1,\cdots$, 并令

$$f(t_j, y(t_j)) - f(t_j, y_j) = L_j e_j,$$

则有

$$\sum_{j=0}^{k} \alpha_j e_{n+j} = h\sum_{j=0}^{k} \beta_j L_j e_{n+j} + h\tau_k(t_n, y(t_n); h). \tag{46}$$

由于 Lipschitz 条件, $|L_j| \leqslant L (j = 0, 1, \cdots)$. 应用引理 5 到(46)式, 取 $z_j = e_j$, $z_{(0)} = e_{(0)}$, $\Lambda = h\tau(h)$, $B^* = BL$ 和 $L^* = B^*\Gamma^*$, 于是得到

$$\| e_n \| \leqslant N_1(\eta) e_{(0)} e^{nhLB\Gamma^*} + \tau(h) \frac{e^{nhLB\Gamma^*} - 1}{BL},$$

其中 $\eta = A\Gamma^* k$, $A = |\alpha_{k-1}| + \cdots + |\alpha_0|$. 令 $t \in [t_0, b]$, $t_n = t_0 + nh = t$, $n \geqslant 0$ 是一个整数, 因为 $e(t;h) = e_n$, 所以推得对 $h \leqslant h_0$, $t - t_0 = nh, n = 0,1,\cdots$, 有

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} e^{LB\Gamma^*(t-t_0)} + \tau(h) \frac{e^{LB\Gamma^*(t-t_0)} - 1}{BL}.$$

证毕.

令线性 k 步法(6)式是 p 阶的, 且假定准确解 $y(t)$对于 $t \in [a, b]$有连续的 $p+1$ 阶导数, 则有

$$\| \tau_k(t, y; h) \| \leqslant Ch^p.$$

因此, (45)式变为

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} e^{LB\Gamma^*(t-t_0)} + Ch^p \frac{e^{LB\Gamma^*(t-t_0)} - 1}{BL}. \tag{47}$$

定理 2 表明我们所需要把变量 t 从界(24)式中项$(t - t_0)\tau(h)$去掉而代之以常数的结果已经达到. 用引理 6 估计(45)式能被再改进一点.

定理 3 在定理 1 的条件下, 如果 $h|\alpha_k^{-1}\beta_k| L < 1, t \in [t_0, b]$, 对于 $h \leqslant h_0, t - t_0 = nh, n = 0,1,\cdots$, 线性 k 步法(6)式的全局离散化误差满足

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} e^{Lb(t-t_0)/c} + \tau(h) \frac{\Gamma(e^{Lb(t-t_0)/c} - 1)}{bL}, \tag{48}$$

其中 $e_{(0)} = \max\limits_{0\leqslant j\leqslant k-1} \| e_j \|$, $\eta = ak/c$, $c = 1 - h|\alpha_k^{-1}\beta_k| L$, a 和 b 由(41)式给出.

现在, 对于一些特殊的方法, 我们需要确定在(47)或(48)式中常数 C, a, b, c 和 Γ 的数值. 对于建立在数值积分基础上的一些特殊方法如显式和隐式 Adams 方法, 有

$$C = C_{p+1} M_{p+1}, \tag{49}$$

其中 C_{p+1}是误差常数(对这些方法 $\sigma_k(1) = 1$). 对所有建立在数值积分上的方法, 因为特征多项式为 $\rho_k(\xi) = \xi^k - \xi^{k-q}$, 其中 $1 \leqslant q \leqslant k$, 所以

$$\frac{1}{1 - \xi^q} = 1 + \xi^q + \xi^{2q} + \cdots.$$

由此可得$|\gamma_l|\leqslant 1$．于是对这些方法 $\Gamma=1$．并且根据(39)和(41)式知对这些方法有 $a=1$．对显式和隐式的 Adams 方法，因为对 $l=0,1,\cdots,\gamma_l=1$，所以表 1 中 b 的数值是容易由(40)式得到．对显式的方法有 $\beta_k=0$，所以 $c=1$．因为对所有基于数值积分的隐式方法 $\alpha_k=1$，所以 $c=1-h|\beta_k|L$.

表 1　Adams 方法中的常数 b

p	1	2	3	4	5	6
显式 Adams 方法 b	1	$\frac{3}{2}$	$\frac{23}{12}$	$\frac{55}{24}$	$\frac{1\,901}{720}$	$\frac{6\,336}{1\,440}$
隐式 Adams 方法 b	1	1	$\frac{13}{12}$	$\frac{28}{24}$	$\frac{897}{720}$	$\frac{1\,902}{1\,440}$

记

$$d=\sup_{(t,y;h)\in S}(\alpha_k^{-1}\beta L(t,y)),\tag{50}$$

其中 $L(t,y)=\frac{\partial}{\partial y}f(t,y)$，利用引理 7 可证得如下结果

定理 4　在定理 1 的条件下，如果 $hd<1,t\in[t_0,b]$，对于 $h\leqslant h_0,t-t_0=nh,n=0,1,\cdots$，线性 k 步法(6)的全局离散化误差满足

$$\|e(t;h)\|\leqslant N_1(\eta)e_{(0)}\mathrm{e}^{Lb(t-t_0)/c}+\tau(h)\frac{\Gamma(\mathrm{e}^{Lb(t-t_0)/c}-1)}{bL},\tag{51}$$

其中 $e_{(0)}=\max\limits_{0\leqslant j\leqslant k-1}\|e_j\|,\eta=ak/c^*,c^*=|1-dh|,a$ 和 b 由(41)式给出，d 由(50)式给出.

2.2　一般 k 步法

在这一部分将给出一般 k 步法(3)的统一的离散化误差估计.

定理 5　对 $t\in[a,b]$考虑有准确解 $y(t)$的初值问题(1)，令函数 $\Phi_k(t,y;h)$满足 Lipschitz 条件，即存在常数 h_0 和 L，使得对所有 $t\in[a,b],0\leqslant h\leqslant h_0,y_j,y_j^*\in\mathbb{R}$ 和 $f\in C^1[a,b]$，

$$\|\Phi(t,y_k,y_{k-1},\cdots,y_0;h)-\Phi(t,y_k^*,y_{k-1}^*,\cdots,y_0^*;h)\|\leqslant L\sum_{j=0}^{k}\|y_j-y_j^*\|,\tag{52}$$

且令相对局部离散化误差

$$\|\tau_k(t,y;h)\|\leqslant\tau(h),\qquad t\in[a,b],h\leqslant h_0,\tag{53}$$

其中 $\tau(h)$依赖于 h 和一些常数，如果(3)式满足相容性和根条件，则对任意的 $h\leqslant h_0$ 和所有 $t\in[t_0,b],a\leqslant t_0,t-t_0=nh,k$ 步法全局离散化误差

$$\|e(t;h)\|\leqslant N_1(\eta)e_{(0)}\mathrm{e}^{kL\Gamma^*(t-t_0)}+\tau(h)\frac{\mathrm{e}^{kL\Gamma^*(t-t_0)}-1}{kL},\tag{54}$$

其中 $e_{(0)}$是由 $e_{(0)}=\max\limits_{0\leqslant j\leqslant k-1}\|e_j\|$ 定义的最大初始值误差，$\eta=A\Gamma^*k,A=|\alpha_{k-1}|+\cdots+|\alpha_0|,\Gamma^*=\Gamma/(1-h|\alpha_k^{-1}|L_k),\Gamma=\sup\limits_{j=0,1,\cdots}|\gamma_j|<\infty,\gamma_j(j=0,1,\cdots)$由(21)式给出，$L_k$ 由下面(59)式定义.

证　由(13)和(14)式可得 $y(t_{n+j})=y(t_n+jh)\ (j=0,1,\cdots,k-1)$ 满足

$$\begin{aligned}&\alpha_{n+k}y(t_{n+k})+\alpha_{n+k-1}y(t_{n+k-1})+\cdots+\alpha_ny(t_n)\\&\quad=h\Phi(t_n,y(t_{n+k}),\cdots,y(t_n);h)+h\tau_{k,n},\end{aligned}\tag{55}$$

其中 $\tau_{k,n}=\tau_k(t_n,y(t_n);h)$．将(55)式减去(3)式，记 $e_{n+j}=y(t_{n+j})-y_{n+j},\gamma_{n+j}=y(t_{n+j};$

$h)(j=0,1,\cdots,k-1)$，我们有

$$\alpha_{n+k}e_{n+k}+\alpha_{n+k-1}e_{n+k-1}+\cdots+\alpha_n e_n$$
$$=h(\Phi(t_n,y(t_{n+k}),\cdots,y(t_n);h)-\Phi(t_n,y_{n+k},\cdots,y_n;h))+h\tau_{k,n}. \quad (56)$$

令

$$\Phi(t_n,y(t_{n+k}),\cdots,y(t_n);h)-\Phi(t_n,y_{n+k},\cdots,y_n;h)$$
$$=\Phi(t_n,y_{n+k}+e_{n+k},\cdots,y_n+e_n;h)-\Phi(t_n,y_{n+k},\cdots,y_n;h)$$
$$=L_{k,n}e_{n+k}+L_{k-1,n}e_{n+k-1}+\cdots+L_{0,n}e_n, \quad (57)$$

利用(57)式，(56)式成为

$$\alpha_{n+k}e_{n+k}+\alpha_{n+k-1}e_{n+k-1}+\cdots+\alpha_n e_n$$
$$=h(L_{k,n}e_{n+k}+L_{k-1,n}e_{n+k-1}+\cdots+L_{0,n}e_n)+h\tau_{k,n}. \quad (58)$$

由于 Lipschitz 条件，$|L_{i,j}|\leqslant L(i=0,1,\cdots,k,j=0,1,\cdots)$. 记

$$L_k=\max_{0\leqslant j\leqslant N}|L_{k,j}|, \quad (59)$$

其中 $N=(b-t_0)/h$，因此应用引理 5，取 $z_j=e_j$，$z_{(0)}=e_{(0)}$，$\Lambda=h\tau(h)$，$B^*=kL$ 和 $L^*=B^*\Gamma^*$，因此得到

$$\|e_n\|\leqslant N_1(\eta)e_{(0)}e^{nhkL\Gamma^*}+\tau(h)\frac{e^{nhkL\Gamma^*}-1}{kL},\quad n=0,1,\cdots,N,$$

令 $t\in[t_0,b]$，$t_n=t_0+nh=t$，$n\geqslant 0$ 是一个整数，因 $e(t;h)=e_n$，故结果对 $h\leqslant h_0$ 和所有 $t\in[t_0,b]$，$a\leqslant t_0$有

$$\|e(t;h)\|\leqslant N_1(\eta)e_{(0)}e^{kL\Gamma^*(t-t_0)}+\tau(h)\frac{e^{kL\Gamma^*(t-t_0)}-1}{kL}.$$

证毕.

显而易见单步法的全局离散化误差先验界和线性多步法的全局离散化误差界(45)式都是界(54)式的特例，就是说界(54)式是所有 k 步法全局离散化误差的统一的先验界. 对多步法，因$k=1$，$A=1$，$L_1=0$，$\Gamma=1$，$\Gamma^*=1$ 和 $N_1(\eta)=1$，所以此时界(54)式和单步法全局离散化误差的经典的先验界相一致. 如果令 $kL=BL$ 或 bL/Γ(注意等号两边的 L 有不同的含义)且 $L_k=|\beta_k|L$，那么界(54)式就与界(45)式相同.

假定方法(3)是 p 阶的，那么我们有

$$\|\tau_k(t,y;h)\|\leqslant Ch^p.$$

由此可得

$$\|e(t;h)\|\leqslant N_1(\eta)e_{(0)}e^{kL\Gamma^*(t-t_0)}+Ch^p\frac{e^{kL\Gamma^*(t-t_0)}-1}{kL}. \quad (60)$$

不同的同阶方法有不同的全局离散化误差界误差，它们是由常数 C 所区分的.

3 改进的累积舍入误差界

为了讨论浮点机上的舍入误差，先来考察一下舍入误差的各个分量. 不失一般性可以假定量 h,t_0,t,α_j 在计算过程中是准确的(即无舍入误差)，则

$$\tilde{y}(t) = \mathrm{fl}(y(t)), \quad t = t_0 + jh, \quad j = 0,1,\cdots,k-1,$$

$$\alpha_k \tilde{y}(t+kh;h) + \mathrm{fl}\Big\{\sum_{j=0}^{k-1}\mathrm{fl}[\alpha_j \tilde{y}(t+jh;h)] - \mathrm{fl}[h\cdot\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))]\Big\} = 0, \quad n = 0,1,\cdots, \tag{61}$$

其中记号 $\mathrm{fl}(x)$代表 x 的浮点舍入值．所以局部舍入误差

$$\varepsilon(t;h) = \Big[\sum_{j=0}^{k-1}\alpha_j \tilde{y}(t+jh;h) - h\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h)\Big] - \mathrm{fl}\Big\{\sum_{j=0}^{k-1}\mathrm{fl}[\alpha_j\tilde{y}(t+jh;h)] - \mathrm{fl}[h\cdot\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))]\Big\},$$

且可以写成如下分量的和：

$$\varepsilon(t;h) = \sigma(t;h) + \pi_j(t;h) + \pi(t;h) + \mu(t;h), \tag{62}$$

其中

$$\sigma(t;h) = \Big\{\sum_{j=0}^{k-1}\mathrm{fl}[\alpha_j\tilde{y}(t+jh;h)] - \mathrm{fl}[h\cdot\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))]\Big\} - \mathrm{fl}\Big\{\sum_{j=0}^{k-1}\mathrm{fl}[\alpha_j\tilde{y}(t+jh;h)] - \mathrm{fl}[h\cdot\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))]\Big\}, \tag{63}$$

$$\pi_j(t;h) = \alpha_j\tilde{y}(t+jh;h) - \mathrm{fl}(\alpha_j\tilde{y}(t+jh;h)), \quad j = 0,1,\cdots,k-1, \tag{64}$$

$$\pi(t;h) = \mathrm{fl}[h\cdot\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))] - [h\cdot\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))], \tag{65}$$

$$\mu(t;h) = h\cdot[\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h)) - \Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h)]. \tag{66}$$

量 $\sigma(t;h)$称为在和 $\sum_{j=0}^{k-1}\mathrm{fl}[\alpha_j\tilde{y}(t+jh;h)] - \mathrm{fl}[h\cdot\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))]$的浮点舍入中的主导(求和)舍入误差，$\pi_j(t;h)$和 $\pi(t;h)$分别是由于乘积 $\alpha_j\tilde{y}(t+jh;h)$和 $h\cdot\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))$的舍入引起的，$\mu(t;h)$是由于函数 Φ 的计算不准确造成的．通常，在实际中步长 h 是如此的小以致于 $\pi(t;h)$和 $\mu(t;h)\ll\sigma(t;h)$，那么就有 $\varepsilon(t;h)\approx\sigma(t;h)+\pi_j(t;h)$．且对于通常的数值方法如所有的单步法和所有基于数值积分的线性多步法，$\pi_j(t;h)=0$，即，局部舍入误差主要是由求和的舍入误差 $\sigma(t;h)$决定的，这也就是$\sigma(t;h)$被称为主导舍入误差的原因．

对于量 h,t_0,t,α_j 在计算过程中不能准确的表示(即它们被舍入到相应的机器数)的情形，也有同样的结论[1)]．

在局部舍入误差满足如下单一性的假设之下：

$$\|\varepsilon_{n+k}\| \leqslant \varepsilon \quad (n = 0,1,\cdots), \tag{67}$$

其中 ε 是一个常数，有经典结果

1) 李建平．微分方程数值积分中的算不准原理及两个普适关系．中国科学院大气物理研究所博士后研究工作报告，1999．174

定理 6[2,3] 令函数 $f(t,y)$对 $t\in[a,b]$是连续和连续可微的且满足 Lipschitz 条件,如果局部舍入误差满足(67)式, 那么累积舍入误差

$$\| r(t;h) \| \leqslant \Gamma^* \left[Akr_{(0)} + (t-t_0)\frac{\varepsilon}{h} \right] e^{L\Gamma^* B(t-t_0)}, \tag{68}$$

其中 $r_{(0)}$是 $r_{(0)} = \max\limits_{0\leqslant j\leqslant k-1} \| r_j \|$ 定义的最大初始舍入误差,$t\in[t_0,b]$,$a\leqslant t_0$,$t-t_0=nh$, 其他常数同定理 1.

先验估计(68)式和传统的全局离散化误差界(27)式有着相同的缺陷. 仿照对全局离散化误差界的改进, 对于线性 k 步法中舍入误差的影响, 可有

定理 7 在定理 6 的条件下, 如果 $h|\alpha_k^{-1}\beta_k|L<1$,$t\in[t_0,b]$, 则对 $h\leqslant h_0$,$t-t_0=nh$, $n=0,1,\cdots$,线性 k 步法(6)的累积舍入误差

$$\| r(t;h) \| \leqslant N_1(\eta) r_{(0)} e^{Lb\Gamma^*(t-t_0)/c} + \frac{\varepsilon}{h}\frac{e^{LB\Gamma^*(t-t_0)}-1}{BL}, \tag{69}$$

其中 $r_{(0)} = \max\limits_{0\leqslant j\leqslant k-1} \| r_j \|$, 其他常数同定理 2.

定理 8 在定理 6 的条件下, 如果 $h|\alpha_k^{-1}\beta_k|L<1$,$t\in[t_0,b]$, 则对 $h\leqslant h_0$,$t-t_0=nh$, $n=0,1,\cdots$,线性 k 步法(6)的累积舍入误差

$$\| r(t;h) \| \leqslant N_1(\eta) r_{(0)} e^{Lb(t-t_0)/c} + \frac{\varepsilon}{h}\frac{\Gamma(e^{Lb(t-t_0)/c}-1)}{bL},$$

其中 $r_{(0)} = \max\limits_{0\leqslant j\leqslant k-1} \| r_j \|$, 其他常数同定理 3.

事实上, (6)式减(61)式,记 $r_j = y_j - \tilde{y}_j$,$j=0,1,\cdots$,并令

$$f(t_j,y_j) - f(t_j,\tilde{y}_j) = L_j r_j,$$

因为 Lipschitz 条件$|L_j|\leqslant L$, 于是得

$$\sum_{j=0}^{k}\alpha_j r_{n+j} = h\sum_{j=0}^{k}\beta_j L_j r_{n+j} - \varepsilon_{n+k}. \tag{70}$$

对这个关系应用引理 6,取 $z_j = r_j$, $\hat{\Lambda} = \varepsilon\Gamma$, $b = bL$, $\eta = ak/c$, $c = 1 - h|\alpha_k^{-1}\beta_k|L$ 和 $z_{(0)} = r_{(0)}$, 可得定理 8.

定理 9 在定理 6 的条件下, 如果 $hd<1$,$t\in[t_0,b]$, 则对 $h\leqslant h_0$,$t-t_0=nh$, $n=0,1$, $\cdots$,线性 k 步法(6)的累积舍入误差

$$\| r(t;h) \| \leqslant N_1(\eta) r_{(0)} e^{Lb(t-t_0)/c^*} + \frac{\varepsilon}{h}\frac{\Gamma(e^{Lb(t-t_0)/c^*}-1)}{bL}, \tag{71}$$

其中 $r_{(0)} = \max\limits_{0\leqslant j\leqslant k-1} \| r_j \|$, 其他常数同定理 4.

对于一般的 k 步法可有如下统一的结果:

定理 10 令 $\Phi_k(t,y;h)$满足(52)式, 如果局部舍入误差满足单一性假设

$$\| \varepsilon(t;h) \| \leqslant \varepsilon, \tag{72}$$

则 k 步法(3)的累积舍入误差

$$\| r(t;h) \| \leqslant N_1(\eta) r_{(0)} e^{kL\Gamma^*(t-t_0)} + \frac{\varepsilon}{h}\frac{e^{kL\Gamma^*(t-t_0)}-1}{kL}, \tag{73}$$

其中 $r_{(0)}$是由 $r_{(0)} = \max\limits_{0\leqslant j\leqslant k-1} \| r_j \|$ 定义的最大初始舍入误差, 其他同定理 5.

显然，界(73)式是界(69)式的推广．包含在界(73)式中的实质结果是累积舍入误差 $r(t;h)$具有 h^{-1}的阶(即与相对局部舍入误差 $\delta=\varepsilon/h$ 有相同的界)，且界(73)式不依赖于对方法的全局离散化误差 $e(t;h)$的界(54)式来说是典型的常数 C 和 p.

上面给出的累积舍入误差界是在单一性假设下得到的，即所有局部舍入误差都是按照它们的最大值的情形进行累积，因此虽然这些界在理论上是有价值的(如对下面的舍入误差概率理论的推导)，但却过高地估计了实际的舍入误差．为了得到符合实际的舍入误差估计，就需要按照舍入误差的“正常”增长来描述．要达到这个目的就必须利用舍入误差的概率理论．Henrici[2,3]非常详细地研究了定点机上舍入误差的概率理论，他的许多结果对浮点机上的情形也同样适用．在对舍入误差进行统计处理之前，必须作出局部舍入误差是随机变量且相互独立的假设，且其分布为 $F(x)$．此外，为了大幅度的简化证明而又不改变最终结果的正确性，对于累积舍入误差 r_n 也假设是独立的(事实上，不作这个假定，也会得到同样的结果，只不过证明过程非常复杂[1])．如前所述，在浮点机上局部舍入误差的重要影响是由主导舍入误差引起的，把其他舍入误差略去，则局部舍入误差 ε_n 的阶为 uy_n，其中 $u=\gamma/10=0.5\times10^{-n}$,这里 γ 为机器精度，n 为有效数字的位数．以 $E(\xi)$，$D(\xi)$分别表示随机变量 ξ 的期望值和方差．容易证明

定理 11　若局部舍入误差 ε_n 是独立的随机变量，则

$$E(\varepsilon_{n+1})=0, \tag{74}$$

$$D(\varepsilon_{n+1})=\frac{1}{3}(uy_n)^2. \tag{75}$$

引理 8　令特征多项式 $\rho_k(\xi)$满足根条件，且系数 $\gamma_j(j=0,1,\cdots)$定义为

$$\frac{1}{\alpha_k^2+\alpha_{k-1}^2\xi+\cdots+\alpha_0^2\xi^k}=\tilde{y}_0+\tilde{y}_1\xi+\tilde{y}_2\xi^2+\cdots, \tag{76}$$

则

$$\tilde{\Gamma}=\sup_{j=0,1,\cdots}|\tilde{\gamma}_j|<\infty.$$

仿照 Henrici 的处理，将舍入误差表示成

$$r_n=\sum_{l=k}^{n}d_{n,l}\varepsilon_l, \tag{77}$$

其中 $r_i=0,i=0,\cdots,k-1,d_{n,l}$是待定常数．对于线性 k 步法，$d_{n,l}$满足

$$\sum_{j=0}^{k}\alpha_j d_{n+j,l}=h\sum_{j=0}^{k}\beta_j L_{n+j}d_{n+j,l},\quad l=k,\cdots,n, \tag{78}$$

$$\sum_{j=J}^{k}\alpha_j d_{n+j,n+j-1}=h\sum_{j=J}^{k}\beta_j L_{n+j}d_{n+j,n+j-1},\quad J=1,\cdots,k-1, \tag{79}$$

$$\alpha_k d_{n+k,n+k}=1+h\beta_k L_{n+k}. \tag{80}$$

根据前面的假设，由(70)式得

$$\sum_{j=0}^{k}\alpha_j^2 D(r_{n+j})=2h\sum_{j=0}^{k}\beta_j L_j D(r_{n+j})+\sigma_{n+1}^2+O(h),$$

1) 见 561 页脚注 1)

其中 $\sigma_{n+1}^2 = D(\varepsilon_{n+1})$. 由 Lipschitz 条件, $|L_j| \leqslant L$, 并应用引理 5 和 $r_i = 0, i = 0, \cdots, k-1$, 得

$$D(r_n) \leqslant (1 + O(h)) \frac{\sigma^2}{h} \frac{e^{2nhL\tilde{\Gamma}^*} - 1}{2BL},$$

其中 $\sigma = \max\limits_{k \leqslant j \leqslant n} \sigma_j$, $\tilde{\Gamma}^* = \tilde{\Gamma}/(1 - h\beta|\alpha_k|^{-1})$. 由此可得

定理 12 对于线性 k 步法(3), 如果局部舍入误差是独立的随机变量, 那么累积舍入误差是期望值为 0 的随机变量, 其方差为

$$D(r(t;h)) \leqslant (1 + O(h)) \frac{\sigma^2}{h} \frac{e^{2LB\tilde{\Gamma}^*(t-t_0)} - 1}{2BL}. \tag{81}$$

仿此可有

定理 13 对于 k 步法(3), 令 $\Phi_k(t, y; h)$ 满足(52)式, 如果局部舍入误差是独立的随机变量, 那么累积舍入误差是期望值为 0 的随机变量, 其方差为

$$D(r(t;h)) \leqslant (1 + O(h)) \frac{\sigma^2}{h} \frac{e^{2kL\tilde{\Gamma}^*(t-t_0)} - 1}{2kL}, \tag{82}$$

其中 $\sigma^2 = \max\limits_{k \leqslant j \leqslant n} D(\varepsilon_j) = \max\limits_{k \leqslant j \leqslant n} (uy_j)^2/3$, $\tilde{\Gamma}^* = \tilde{\Gamma}/(1 - h|\alpha_k|^{-1}L_k)$, L_k 同定理 5.

这个结果表明, 舍入误差的"正常"累积增长是由随机变量 $r(t;h)$ 的标准差所表征的, 其阶为 $h^{-1/2}$, 比前面在单一性假设下所得的理论上界改进了 $h^{1/2}$ 的因子. 在下面的分析中, 将略去高阶小项 $O(h)$. 这样由概率理论所得的累积舍入误差的界为

$$\| r(t;h) \| \sim \sigma \frac{e^{kL\tilde{\Gamma}^*(t-t_0)}}{\sqrt{2hkL}}.$$

4 计算不确定性原理

有了前面的误差估计, 我们就可以解释在文献[1]中所观察到的各种数值现象. 根据前面的结果, 一个 p 阶的 k 步法(3)当初值是完全准确的且无初始舍入误差时, 易知其总误差满足

$$\| E(t;h) \| = \| e(t;h) + r(t;h) \| \leqslant C(t)\tilde{E}(h) = C(t)\left(Ch^p + \frac{\sigma}{\tilde{C}\sqrt{h}} \right), \tag{83}$$

其中 $C(t) = e^{C_L\hat{\Gamma}(t-t_0)}/\sqrt{C_L}$ 为时间函数, $\hat{\Gamma} = \max(\Gamma^*, \tilde{\Gamma}^*)$, $\tilde{C} = \sqrt{2C_L}$, C 是与方法有关的常数, C_L 是与微分方程有关的常数, $\sigma = \max\limits_{t \in [t_0, t_n]} u \| y(t) \| /\sqrt{3}$, $u = \gamma/10 = 0.5 \times 10^{-n}$, 这里 γ 为机器精度, n 为有效数字的位数. 对于单步法, $C_L = L$. 对于线性多步法, $C_L = BL$ 或 bL/Γ. 对于一般 k 步法, $C_L = kL$. 对于 Taylor 级数法、显式和隐式 Adams 方法 $C = C_{p+1}M_{p+1}$, C_{p+1} 为误差常数, $M_{p+1} = \max\limits_{t \in [t_0, t_n]} \| y^{(p+1)}(t) \|$. 易证

定理 14 当

$$h = H = \left(\frac{\sigma}{2pC\tilde{C}} \right)^{1/(p+0.5)} \tag{84}$$

时, $\tilde{E}(h)$ 达到最小值.

这表明由于机器的有限精度, 所以存在一个最优步长 H, 除此而外, 误差不可能再得到改进. 这就是在文献[1]数值试验中所观察到的存在最优步长的原因. 当提高机器精度时,

舍入误差就会减少，所以由(84)式知，对应的最优步长也要相应地缩小；当 C 越小时，H 越大；p 越大，H 越大．这解释了文献[1]中所得到的最优步长随方法的阶数的增加而增加、在双精度下比单精度的小、Taylor 级数法和隐式 Adams 型方法比同阶的显式 Adams 法大的结果．此外，如果 $y(t)\in C^{\infty}[a,b](t\in[a,b])$，那么由(84)式知当 $p\to\infty$ 时 $H\to 1$.

定理 15　用同一种 p 阶的 k 步法分别在具有 n_1,n_2 位有效数字的两种机器精度 γ_1,γ_2 $(\gamma_1=5\times10^{-n_1},\gamma_2=5\times10^{-n_2},n_1\leqslant n_2)$下对一个微分方程进行积分，则它们的最优步长 H_1，H_2 之比为

$$l=\frac{H_1}{H_2}=10^{\frac{\Delta n}{p+0.5}},\tag{85}$$

其中 $\Delta n=n_2-n_1$.

这证实了文献[1]中发现的 l 所满足的普适关系．这个关系表明 l 只与方法的阶数和机器的精度(即有效数字的位数)有关，而与方程的类型、初值及方法本身无关．所以，只要知道了某一机器精度下的最优步长，那么由它立得任何机器精度下的最优步长．若给定 n_1,n_2，则当 $p\to\infty$ 时 $l\to 1$.

根据上面的结果知，当机器精度给定时，数值方法所得数值解能达到的最好准确度就被完全确定，而这个最好准确度与最优步长相对应．由(4)式可得最优步长下的误差公式为

$$\|E(t;H)\|\approx C(t)\frac{\sigma}{\tilde{C}\sqrt{H}}\left(1+\frac{1}{2p}\right).\tag{86}$$

这表明误差随 H 或 p 的增大而减小，随 σ(即 γ)的减小而减小．如果给定误差容限 $\delta>0$，由(86)式确定的达到这个误差容限所需的积分时间即为最大有效计算时间 $T(T=t-t_0)$，即

$$C(T)=\frac{\delta\tilde{C}\sqrt{H}}{\sigma(1+1/2p)}.\tag{87}$$

显然，最大有效计算时间 T 随 H 或 p 的增大而增大，而随 σ(即 γ)的减小而增大．这解释了文献[1]中的数值结果：最大有效计算时间在双精度下比单精度的长，高阶方法较低阶方法长，RK 型、Taylor 级数型和隐式 Adams 型方法的比同阶的显式 Adams 型方法略长．

定理 16　用同一种 p 阶的 k 步法在两种机器精度 $\gamma_1,\gamma_2(\gamma_1\geqslant\gamma_2)$下对一个微分方程进行积分，则它们的最大有效计算时间函数 $C(T_1),C(T_2)$ 之比为

$$k=\frac{C(T_2)}{C(T_1)}=l^p.\tag{88}$$

这也是一个普适关系．由这个关系可以推得

$$\Delta T=\hat{C}\cdot p\ln l,\tag{89}$$

其中 $\Delta T=T_2-T_1,\hat{C}=(C_l\hat{\Gamma})^{-1}$．当 $p\to\infty$ 时，$k\to\gamma_2/\gamma_1=10^{\Delta n}$，$\lim\limits_{p\to\infty}\hat{C}^{-1}\Delta T=\Delta n\ln 10$．在文献[1]中的两种精度下$(n_1=7,n_2=16,\Delta n=9)$，$\lim\limits_{p\to\infty}\hat{C}^{-1}\Delta T=9\ln 10$．这就解释了文献[1]中得到的双精度和单精度下的最大有效计算时间的差随着阶数的增加趋近于一定值的结论．

上述理论分析指出，当考虑了舍入误差后，在文献[1]数值试验中所发现的现象能被很好地解释．进一步，在此基础上，给出文献[1]中所提出的计算不确定性原理的数学表述．为此将 $\tilde{E}$ 表述成 $\tilde{E}=\tilde{e}+\tilde{r}$，其中 $\tilde{e}=Ch^p,\tilde{r}=\sigma/\tilde{C}\sqrt{h}$，并将 σ 用机器精度 γ 表成 $\sigma=C_\sigma\gamma$，这

里 $C_\sigma = \max\limits_{t\in[t_0,t_n]} \| y(t)\| /10\sqrt{3}$为常数. $\tilde{e}$(实质上表征了整体离散误差 $e(t;h)$的本质部分)是数值方法的不准确所带来的不确定性的度量, $\tilde{r}$(实质上表征了累积舍入误差 $r(t;h)$的本质部分)则是机器的有限精度所造成的不确定性的度量, 而 $\tilde{E}$ 是这两种不确定性之和.

定理 17 只要机器精度是有限的, 那么 $\tilde{E}$ 就不会趋于零, 即

$$\tilde{e} + \tilde{r} \geqslant C_{\min}, \tag{90}$$

其中 $C_{\min} = (1+2p)[C(C_\sigma\gamma/2p\tilde{C})^{-2p}]^{1/(2p+1)}$.

这个结果说明无论步长如何的小, 总误差都不可能任意的小, 除非机器精度 $\gamma\to 0$. 当不考虑舍入误差时, 全局离散化误差随步长趋于零而趋于零, 因此, 此时数值解是理论收敛的. 然而, 在实践中, 由于机器的有限精度所导致的舍入误差是不可避免地存在的, 使得一开始当步长减小时, 全局离散化误差减小, 总的误差也减小; 然后, 当步长进一步减小时, 舍入误差却愈来愈大, 总的误差又开始增加(总误差由减小变为增加的转折点即对应于最优步长). 所以, 当 $h\to 0$ 时, 数值解是理论收敛的, 但如果用有限精度计算的话, 那么它不是数值收敛的. 换句话说, 在有限机器精度下, 数值收敛性和理论收敛性是不能同时发生的, 即, 在实际中当 $h\to 0$ 时数值解不是真实收敛的. 为了得到更高精度的数值解, 随着步长的减小就不得不增加机器的精度. 进一步, 有

定理 18 令 $\tilde{e}^* = \tilde{e}^{1/2p}$, 则

$$\tilde{e}^* \cdot \tilde{r} = \hbar_p, \tag{91}$$

其中 $\hbar_p = \gamma C_\sigma C^{1/2p}/\tilde{C}$.

此即计算不确定性原理, 这是量子力学中著名的测不准关系[11, 12]在数值计算中的表现. 它表明由数值方法所导致的不确定性和由计算机所决定的不确定性是两个"共轭"量, 它们不可能同时减小到零, 若其中之一的不确定性越小, 则它的"共轭"量的不确定性就越大. 由于这两种不确定性之间存在的这种固有关系, 就使得数值解的有效区间长度受到限制, 这是造成最大有效计算时间必然存在的根源. 就是说, 若给定误差容限 $\delta > 0$(即误差小于这个容限的数值解是可接受的), 那么必然存在最大有效计算时间 T, 在区间$[0,T]$内数值解满足这个容限的要求并把真解较好的再现出来, 而在这个区间之外的真解是数值方法无法确定的. 因此, 在有限计算精度下, 计算不确定性原理给数值解法的计算能力加上了确定的限制.

5 结论与讨论

本文在常微分方程数值解法误差分析的经典结果的基础上, 研究了常微分方程一般数值解法的误差传播规律, 指出经典结果的缺陷. 根据所讨论问题的性质和特点, 除了传统的收敛性概念(即本文中的理论收敛性)外, 又提出两种新的收敛性概念: 数值收敛性和真实收敛性, 并对浮点机上一般数值解法的舍入误差的各种分量作了详细讨论. 通过引进一类新的递推不等式和利用概率理论, 不仅本质改进了线性多步法误差界的经典结果, 导出了浮点机上舍入误差的"正常"累积增长, 而且给出一般多步法总误差的统一估计. 根据所得误差分析的结果, 解释了文献[1]数值试验中所观察到的各种现象, 导得两个与方程、初值、数值格式无关的普适关系, 它们与数值试验中的结果相一致, 并指出在实际中, 数值解的理论收敛性和数值收敛性是不能同时发生的, 即它不是真实收敛的. 进一步的理论分析给出计算不确定性

原理的明确数学表述，由此阐明了数值解法和计算机所带来的两种不确定性是两个“共轭”量，它们不可能同时减小到零，从而解释了数值方法在有限机器精度下计算必然存在最优步长和最大有效计算时间的根源．由计算不确定性原理知，为了获得更高准确性的和更长有效范围的数值解，就必须提高计算机的精度．

此外，利用本文的结果，对于用计算机进行数值模拟和预测的问题还有如下一些附加讨论：(1) 在现实当中，计算不确定性原理指出计算机的有效模拟能力是有限度的．我们必须认识到这一点．有效模拟能力之所以会有极限是因为除去一个零测度集外，计算误差是完全不可避免的．这个界限的存在性是固有的，是不依赖于所模拟的对象(确切的说是除去一个零测度集)的，而这个界限的大小常常是与模拟的对象有关．一旦研究的对象和计算机的精度给定，那么所能模拟的最好能力便被确定下来，这个限制同样是固有的，是不能通过改进描述这个对象的模式或改善资料来加以克服的；而改进描述这个对象的模式或改善资料只能是使模拟的能力逐渐接近这个最好程度．(2) 利用计算不确定性原理使模拟达到最好程度．计算不确定性原理一方面指出了模拟能力的限度，另一方面又指出了最优关系．这个最优关系给出达到最好模拟能力的途径．根据这个关系，我们必须确认哪些计算结果是有效的、可以肯定的，哪些结果是无效的、不能确定的，从而明确数值预测结果中的正确部分．(3) 发展高精度的计算机是提高有效计算能力的一条途径．目前对于微分方程的各种数值解法，其核心都是步进式的递推过程，而这种方法必然存在最大有效积分时间，超过这个时间的积分结果将是无效的，从而不能对系统长期行为作出正确的分析．而根据计算不确定性原理，只要增加机器精度就可以延长最大有效计算时间，从而提高有效计算能力．总之，目前正面临着从无限精度理想化向有限精度现实性的观念转变，在这个转变过程中，如何冲破在有限机器精度下的计算不确定性原理并提高长时间数值计算能力则是需要解决的重要课题．

参　考　文　献

1 李建平，曹庆存，丑纪范．非线性常微分方程的计算不确定性原理——Ⅰ．数值结果．中国科学，E 辑，2000，30(5)：403～412

2 Henrici P. Discrete Variable Methods in Ordinary Differential Equations. New York: John Wiley, 1962. 1～165, 187～288

3 Henrici P. Error Propagation for Difference Methods. New York: John Whiley and Sons, 1963

4 Gear C W. Numerical Initial Value Problems in Ordinary Differential Equations. Englewood Cliffs Prentice-Hall, 1971. 1～14, 72～86

5 Hairer E, Nrsett S P, Wanner G. Solving Ordinary Differential Equations I. Nonstiff Problems. 2 nd ed. Berlin, Heidelberg, New York: Springer-Verlag, 1993. 130～430

6 Stoer J, Bulirsch R. Introduction to Numerical Analysis. 2nd ed. Berlin, Heidelberg, New York: Springer-Verlag, 1998. 1～36, 428～569

7 李庆扬．常微分方程数值解法(刚性问题与边值问题)．北京：高等教育出版社，1991

8 李荣华，冯国枕．微分方程数值解法．北京：高等教育出版社，1990. 1～64

9 Dahlquist G. Convergence and stability in the numerical integration of ordinary differential equations. Math Scandinavica, 1956, 4: 33～53

10 Dahlquist G. 33 years of numerical instability, Part I. BIT, 1985, 25: 188～204

11 Heisenberg W. The Physical Principles of Quantum Theory. Chicago: University of Chicago Press, 1930

12 McMurry S M. Quantum Mechanics. London: Addison-Wesley Longman Ltd, 1998